View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by PhilPapers

(Online) Manipulation: Sometimes Hidden, Always Careless

Michael Klenk

Delft University of Technology

m.b.o.t.klenk@tudelft.nl

Author's manuscript version. This article has been accepted for publication in

REVIEW OF SOCIAL ECONOMY. Please cite final published version.

Abstract

Ever-increasing numbers of human interactions with intelligent software agents, online and offline, and their increasing ability to influence humans have prompted a surge in attention toward the concept of (online) manipulation. Several scholars have argued that manipulative influence is always hidden. But manipulation is sometimes overt, and when this is acknowledged the distinction between manipulation and other forms of social influence becomes problematic. Therefore, we need a better conceptualisation of manipulation that allows it to be overt and yet clearly distinct from related concepts of social influence. I argue that manipulation is careless influence, show how this account helps to alleviate the shortcomings of the hidden influence view of manipulation, and derive implications for digital ethics.

Keywords

Manipulation; social influence; autonomy; digital ethics; social expectations.

Acknowledgements

Many thanks to Lisa Herzog and Verina Wild for hosting a stimulating workshop on 'Digital Behavioural Technology' at TU Munich in 2019, which is where many ideas for this paper originated. I am also grateful for comments from participants at the workshop, Thomas Beschorner, Fleur Jongepier, and two anonymous reviewers and editors of this journal. My work on this paper has been supported by a Niels Stensen Fellowship.

***

Introduction

We increasingly interact with intelligent machines, online and offline (Floridi 2014). According to several scholars, many of those interactions are manipulative (e.g. Frischmann and Selinger 2018; Susser, Roessler, and Nissenbaum 2019b; Zuboff 2019; Klenk In press). This critical perspective on interactions with intelligent machines is part of a wider social and scholarly development that views human-machine interactions as often deeply problematic instances of "techno-social engineering", that is, as "processes where technologies and social forces align and impact how we think, perceive, and act" in worrisome ways (Frischmann and Selinger 2018, 4). Going beyond concerns about lack of fairness in exchanges, human interactions with intelligent machines appear to be of a much more worrisome nature: they seem deeply *manipulative*, with drastic consequences for human well-being and autonomy (Susser, Roessler, and Nissenbaum 2019b).

Questions about manipulation occupy an important place in the debate about the ethics of (digital) behavioural technologies. Behavioural technologies are technologies designed to change, or support change in, human behaviour. For example, the health app 'Headspace' aims to support users with meditations and the shift toward a more mindful lifestyle. Behavioural technologies raise ethical questions about the legitimacy of the behaviour they aim to support, but also, and perhaps more importantly, about the legitimacy of the type of influence they have on users. Digital behavioural technologies – with their comparatively cheap 'deployment' costs, wide availability, and potential to optimise their means of influencing by learning from data – only exacerbate the urgency of these questions.

Questions about manipulation are important for this debate in at least two ways. First, digital behavioural technologies can be studied as *tools* wielded by humans or firms, and questions about manipulation would concern whether these tools are used in a manipulative way. Second, digital behaviour technologies – which sometimes operate autonomously (such as a recommender system designed to keep users engaged) – may themselves be considered as agents of manipulation. This may be the case because they embody ethical characteristics (Klenk 2020), or because they are indeed agents in a relevant sense. In any case, it can be asked whether digital behavioural technologies are themselves manipulative or whether they are used as tools in a manipulative way.

In light of the controversy surrounding our interactions with machines and the relevance of manipulation for the ethics of behavioural technologies, it becomes crucial to understand what manipulation is, so that problematic types of interactions can be distinguished from unproblematic ones. In the popular understanding of manipulation (e.g. Merriam-Webster 2020), and several recent contributions to the debate in digital ethics from a philosophical (Susser, Roessler, and Nissenbaum 2019b, 2019a) and legal perspective (Zarsky 2019), the view is that covertness is a necessary component of manipulation. Call this the covertness thesis:

Covertness thesis: Manipulation is hidden influence.

However, despite its intuitive appeal, prominence, and influence, the covertness thesis is false. Hidden influence is not a necessary component of

manipulation.[1] One possible counterexample against the covertness thesis is a manipulative guilt trip, and there are many others (as I will show below). Yet, rejecting the covertness thesis leads to a hitherto unnoticed problem about demarcating manipulation from other forms of social influence, notably coercion. Therefore, the challenge is to rectify the conception of manipulation used in digital ethics, while maintaining a useful distinction from other forms of social influence.

In this paper, I explain why the covertness thesis is initially attractive but ultimately false, in a way that goes beyond counterexamples by raising the novel *demarcation problem*. I then demonstrate how to overcome the demarcation problem by introducing a suitable account of manipulative action as a form of *careless influence*. The paper thus provides a deeper understanding of manipulation and its application to digital ethics and should be of interest to philosophers, economists, and legal scholars studying the ethical implications of digitalisation.

I proceed as follows. First, I introduce the covertness thesis in more detail. I partly cover familiar territory here, but I also add something new by showing how the covertness thesis helps to avert a dilemma about manipulation. Then I show why the covertness thesis fails. I use counterexamples as well as deeper theoretical and practical arguments. Finally, I propose a way out of the problem by arguing that manipulative action is always *careless influence*.

[1] Neither is it a sufficient one. For example, consider that you might simply fail to attend to someone's influence on you, even though your interlocutor does not aim to hide their influence from you. Such cases (e.g. a nurse assisting in an operation on you) are not manipulative. I will not focus on the sufficiency of covertness for manipulation in what follows, because its proponents present the covertness thesis as describing a necessary condition.

1 Thesis: Manipulation as hidden influence

According to a popular and increasingly influential view, manipulation is a type of social influence that is necessarily covert. We will see in this section that the covertness thesis implies that *some form of deception is necessary for manipulation*. Recall, the covertness thesis is as follows:

Covertness thesis: Manipulation is hidden influence.

A normatively more loaded formulation of the covertness thesis would hold that manipulation is trickery, or that it always involves cunning, a ruse, ploy, or subterfuge. As all of these types of interpersonal interaction at bottom involve hidden influence, I leave it to the reader to interpret the covertness thesis in his or her favourite terms.[2]

The covertness thesis has been defended by several scholars, in the literature in normative ethics and digital ethics. In an early conceptual analysis of manipulation, Joel Rudinow (1978) suggests that manipulation is a special form of social influence precisely because it is hidden (1978, 339):

> Both persuasion and coercion seem to be open in seeking to influence behavior, whereas manipulation seems in some way more mysterious about it.

Rudinow does not go on to provide a more explicit defence of the covertness of manipulation since his interests lie elsewhere. Like Rudinow, several other normative ethicists hint at the link between manipulation and covert influence. Teun van Dijk defines manipulation as a process of social influence that "happens

[2] Of course, on the covertness thesis, manipulation is not *just* hidden influence. Additional (necessary and sufficient) factors account for some influence being manipulative. But these additional factors shall not concern us in what follows as I will focus exclusively on covertness as an alleged necessary requirement of manipulation.

without the awareness of the recipients" (van Dijk 1998, 276). Robert Goodin (1980, 9) attests that manipulation contains an element of deviousness. Marcia Baron suggests, in a somewhat weaker claim, that manipulation "typically" works best if hidden and that the manipulator, if at all possible, aims to keep the manipulation hidden from the manipulatee (Baron 2003, 39). Alan Ware hints at the covertness thesis to distinguish manipulation from education (Ware 1981, 171). In recent discussions of manipulation in legal contexts, the covertness thesis has also often played a prominent role (Sunstein 2016a; Zarsky 2019). Even though the endorsement of the covertness thesis remains somewhat implicit, I take the views cited above as evidence that there is much sympathy for the view represented by the covertness thesis in discussions of manipulation in normative ethics.

In digital ethics, the covertness thesis has been defended much more explicitly and forcefully, notably by Susser, Roessler, and Nissenbaum (2019a, 2019b). In what follows, I will primarily draw on their account to flesh out the details and implications of the covertness thesis. Susser et al. write that "manipulation is *hidden influence*." They express the view that *covertly* influencing someone is manipulation (Susser, Roessler, and Nissenbaum 2019b, 4). They further specify (Susser, Roessler, and Nissenbaum 2019b, 4):

> Covertly influencing someone—imposing a hidden influence—means influencing them in a way they aren't consciously aware of, and in a way they couldn't easily become aware of were they to try and understand what was impacting their decision-making process.

The view could be either that manipulation requires successfully making it the case that one's influence remains hidden, or merely intending it to be so. The

difference is one of success: in analogy, I can successfully make it the case that you are happy, or merely intend you to be happy. Susser et al. go for the former, more demanding option. They specify as follows (Susser, Roessler, and Nissenbaum 2019b, 5):

> Strictly speaking we would only say that they have succeeded in manipulating when their target is unaware of their machinations.

Thus, the proper interpretation of the covertness thesis is that manipulation is defined by successfully hidden influence. For example, the Trojan horse scheme will count as a case of manipulation, partly because it turns out to be successful. Had the scheme failed, it would have been an attempt only and would thus not count as manipulation, according to the covertness thesis.[3]

The covertness thesis carries an intriguing and, thus far, largely unnoticed connection to another prominent view of manipulation, the view that manipulation always involves deception, or the manipulation-as-deception view for short (cf. Gorin 2014, 74ff. for critical discussion). The covertness thesis and the manipulation-as-deception view are not obviously equivalent because 'deception' is itself a loaded normative concept, which is sometimes interpreted narrowly in aetiological terms. In that case, hidden influences do not per se constitute deception (that is, a hidden influence need not lead to false beliefs in the manipulatee). But there is a broader and more plausible interpretation of deception which shows that the covertness thesis indeed implies the

[3] Of course, this leaves open the possibility that *un*intentionally but covertly influencing someone can count as manipulation. But that does not seem right for reasons that are beyond the scope of this paper, and so it seems that intentionally hiding one's influence *and* succeeding is what makes for manipulation. For reasons of space, I shall not pursue this issue in what follows.

manipulation-as-deception view. Let us then trace the connection from the covertness to the deception view.

Note first that the covertness thesis implies that manipulation requires deception, in the sense that the manipulator deceives about his or her aims or means of influence, which raises the following question: Does deceiving about one's aims or means of influence count as deception? Plausibly, the answer is yes. Manipulation, according to the covertness thesis, requires the manipulatee to be lead astray about the intentions and means of influence of the manipulator. But leading someone astray about one's intentions is deceptive, on a plausible interpretation of 'deception' and so the covertness thesis entails the manipulation-as-deception view.[4]

Therefore, if the covertness thesis is true, then there cannot be manipulation without deception, broadly construed. Manipulation in these cases relies on the fact that subjects form no attitude or a false attitude about the manipulator's intentions and that plausibly counts as deception (Fallis 2009).[5] Being clear about the entailment will be important when assessing the covertness thesis. As we will see below, insofar as there can be non-deceptive manipulation (broadly construed), the covertness thesis must fail.

Though I will later attack the covertness thesis, it is important to emphasise its attractiveness. First and foremost, it seems to capture very well the ordinary

[4] Of course, deception narrowly construed as intentionally causing false beliefs in one's victim may thus be a proper subset of manipulation, and so there can be manipulation without deception narrowly construed (cf. Susser, Roessler, and Nissenbaum 2019b; Cohen 2018; Krstić and Saville 2019).

[5] (Gorin 2014, 79) defends a similar conclusion in reference to a "transparency norm" that requires "that an interactive partner not hide her intentions in interacting when these intentions are relevant to the intentions or interests of the person with whom she is interacting."

user of the term of manipulation, which makes it intuitively compelling.[6] More importantly, it is theoretically useful in distinguishing manipulation from other forms of social influence, notably coercion and persuasion. In contrast to persuasion and coercion, which Susser et al. consider to be "forthright concepts", they claim that "manipulation is hidden – we only learn that someone was trying to steer our decision-making after the fact if we ever find out at all" (Susser, Roessler, and Nissenbaum 2019b, 4).

It helps to pause here for a moment for a digression on broader normative methodology. When we define manipulation, we are partly charting the field of social influence, specifically by setting manipulation in relation to other types of social influence like persuasion and coercion (and, of course, there is the option that some turn out to be just the same). The covertness thesis allegedly has the strength of helping us succeed at this task because it gives us a good criterion to distinguish manipulation from other types of social influence. This task of distinction occupies all analysts of manipulation (cf. Ware 1981, 171).

The contrast that seems most relevant in drawing the boundary of manipulation is between manipulation and coercion. It is commonly assumed that manipulated people are still acting autonomously or out of their free will, even though their agency may be somewhat diminished because of the manipulative influence (cf. van Dijk 1998, 274; Wood 2014). But the assumption of a continuum of autonomy in the victim between manipulation (some agency) and coercion

[6] Most compellingly, it invokes the image of the puppeteer who pulls the strings of his puppet in the background.

(much less or even no agency) raises a deep puzzle: Where can we legitimately and firmly draw a line between both concepts, if not through the criterion of agency?

The covertness thesis helps us answer that question. Indeed, the most significant factor in favour of the covertness thesis is that it helps us to avoid what I call the *demarcation dilemma for theories of manipulation*. Suppose coercion involves curtailing the agency or autonomy of the coerced person. That assumption gains plausibility from the fact that coerced people are usually, at least partly, exculpated for their actions because they did not act in the proper sense (cf. Wood 2014). Now, manipulation also involves influence and, more than that, a degree of control by the manipulator, so that the manipulatee is steered towards the manipulator's aims. There is thus a certain sense in which manipulation also seems to involve undermining the manipulatee's agency. That is aptly captured in the popular image of a puppet master pulling the strings (or pushing the emotional buttons) of his or her unsuspecting victims, leading to determined outcomes. However, if the influence were not successfully hidden, then it would seem that the manipulatee would *willingly* engage in the behaviour. This would raise the question of explaining what is wrong with that behaviour. If the manipulatee does *not* willingly engage in the behaviour, then the manipulatee is not in the strict sense acting at all, and we would be talking about coercion, the annihilation of agency, and not manipulation.[7] So, manipulation might just as well disappear as a phenomenon in the wake of the demarcation dilemma.

[7] More precisely, coercion annihilates choice and deprives the coerced person of her self-chosen ends. In a limited sense, however, she still acts in giving in to coercion. I thank [omitted for blind review] for discussion on this point. Manipulated behaviour, on my view, is less free behaviour, though not necessarily less agentive behaviour. The defence of that view must wait for another paper, however.

The covertness thesis offers a helpful way out by disallowing that overt influences count as manipulation. Therefore, the covertness thesis seems to be in good standing theoretically, not only because of its intuitive plausibility but also because of the important and straightforward solution to the demarcation dilemma it provides.

There are practical benefits to the covertness thesis, too. The covertness thesis explains well why many journalists, academics, and members of the wider public are now concerned with manipulation in the wake of online technologies: there seem to be new vulnerabilities and power hierarchies such that *hidden influence* is much more likely in this context and, thus, there is fertile ground for manipulation. Indeed, as commentators have documented at great length, the potential and actual degree of knowledge over human users and subsequent influence exerted on human users by intelligent machines is quite dramatic (Kramer, Guillory, and Hancock, J. 2014; Burr and Cristianini 2019; Burr, Cristianini, and Ladyman 2018). Crucially, what critics often highlight is the *subtlety* and *covertness* with which we are influenced online. For example, in their discussion of how technology "re-engineers" humans, Frischmann and Selinger (2018) put much emphasis on the fact that much of it happens subtly, aided by now ubiquitous surveillance technologies.

These are important observations because they point to the practical significance of the covertness thesis: people are concerned about hidden influence. That might be a surface-level effect of deeper concerns about autonomy and rational decision-making (both of which, arguably, are hampered by hidden influences, though see Klenk and Hancock, J. 2019). Nonetheless, starting from

the view that there is a problem with online influence, the covertness thesis partially explains and indeed justifies the recent concern with digital technologies.

With the covertness thesis in place, we could focus our efforts on identifying the other components of manipulation, with the required policy recommendations following suit. However, as I will show next, the covertness thesis fails.

2 Antithesis: Overt manipulation

So far, I have introduced the covertness thesis of manipulation, connected it to the important manipulation-as-deception view, and emphasised its theoretical advantages. Rather than merely covering familiar territory, I have highlighted an often neglected theoretical advantage of the Covertness view, which is that it avoids the demarcation dilemma for theories of manipulation. But in this section, I will show that the covertness thesis cannot be correct. First, I will show that there are counterexamples to it. Second, I will show that the covertness thesis has problematic implications that speak against it independently of whether one accepts the relevancy of the offered counterexamples or not.

Some manipulation is overt. For example, consider the following case from Barnhill (2014, 60):

> Janice has booked a vacation trip to New York City. Janice's father, Mike, doesn't want her to go because he thinks that New York City is too dangerous a place. Over the course of a weekend together, Mike repeatedly says things like, "If you go, your mother and I will be sick with fear!" knowing that this will make Janice feel extremely guilty. Mike thinks that it's appropriate for Janice to feel extremely guilty for making her parents worry so much. This makes Janice feel very guilty and as a result, she cancels the trip.

Such manipulative guilt trips can be "plain as the day," according to Barnhill (2014, 60).[8] Hence, hidden influence is not necessary for manipulation. One might resist this conclusion by pointing out that Barnhill's case does not clearly involve harm done to the manipulatee, and so fails to qualify as manipulation. But consider the following case from Gorin (2014, 80–81):

> Wilson and Adams are up for promotion, though only one of them will get the job. Wilson is a recovering alcoholic and Adams sets out to encourage a relapse, intending this to disqualify Wilson for the promotion. Adams consistently drinks alcohol in front of Wilson, offers her alcoholic beverages, vividly describes to her whatever benefits there are to drinking and to drunkenness, and so on, all the while making no secret of his intentions. During a moment of weakness brought on by a particularly difficult and stressful event Adams takes a drink, which leads to more drinks, missed days at work, and an overall decreased ability to meet the demands of her job. When the time comes to announce who will be promoted, Adams is told by her managers that her recent poor performance has made it impossible for them to give her the new job and that they have selected Wilson for the promotion.

Clearly, Adam's behaviour is manipulative, but it does not seem to involve deception or hidden influence at all. An analogous case from the digital realm illustrates the point even more convincingly, given the ubiquity of comparable situations:

> Zaire works in online marketing and advises clients on social media marketing strategies. She is well aware and indeed an expert in online behavioural targeting. As her 30th birthday approaches, she sees more and more targeted advertisements for baby products online. So far, she did not plan on having kids and did not concern herself with the topic. But although she does her best to ignore this obvious marketing effort, she begins to doubt that she should not have kids.

[8] We can even imagine that Janice is very well aware why Mike puts up a such a show – it might be common knowledge between them, or Mike might at some point even admit his ulterior motives. Janice would likely react as she does all the same, especially if Mike is also sincere about his act. Thanks to [omitted for review] for this suggestion.

The marketing efforts seem manipulative, despite Zaire being well aware of the underlying process of influence. The ulterior intention behind the marketeer's influence, a profit motive, is plain as the day to Zaire and yet she seems manipulated (that fact about ulterior motives might have been obscured in earlier cases, but it isn't in this case). Indeed, the fact that the ads prove effective *despite* her being aware of it seems particularly pernicious about this example (a point to which I will return below). Is that perhaps because these two cases involve affective influence, one that the victim is not properly aware of after all? Even if that were the case, that explanation is not available for the following two cases:

> During World War II, American soldiers would fight only if it was common knowledge that other soldiers would not defect. Knowing that the American soldiers were racially divided, Japanese propagandists claimed that all African American (AA) defectors will be well treated. Although the Japanese realized that the AA soldiers knew both that they would not be well treated and that this was just a trick meant to plant distrust within their ranks, the Japanese hoped that white soldiers would not know that the AA soldiers knew this. And even if none of the white soldiers was racist, they would believe that some of their white comrades were racists. The mere spectre of racism was enough to undermine the common knowledge that AA soldiers would not defect. As a result of this general distrust, the AA units were withdrawn, and the Japanese succeeded in reducing American troop strength.

In this case by Krstić and Saville (2019), the Americans were manipulated, despite the fact that they were not deceived by the Japanese and that they were aware of the scheme in terms of its affective influence (concretely, that there was an appeal to racist tendencies). Going digital again makes it straightforward to find a suitable case from everyday experience:

> Marianne is an avid Twitter user and she has a 'folk theory' about the functioning of the algorithm that curates her Twitter feed. Like many others (cf. DeVito et al. 2018), she believes that the algorithm will prioritise and show content to her that is emotionally activating and that appeals to problematic affective tendencies like in-group favouritism. She's a moderate

> Republican and well aware that the more radical people she follows on Twitter aim to push her emotional buttons. Still, she can't help but feel heightened resentment toward the media and the Left after browsing Twitter.

Marianne is not deceived about the intentions of the right-wing radicals on Twitter. Yet, she seems manipulated.[9] In all five cases, we can observe social influence that is plausibly characterised as manipulation even though the influence is by no means hidden. Hence, these are counterexamples to the covertness thesis and, therefore, the covertness thesis must be false.

It is methodologically undesirable to bank one's case entirely on counterexamples that appeal to intuition. But there is more support for the case against the covertness thesis. Some authors like Wood (2014) have recorded a particular moral aversion against *overt* manipulation. In a discussion of manipulative advertising, which Wood argues often involves no deception, Wood finds the "fact that we are manipulated even without actually being deceived more appalling than reassuring" (Wood 2014, 38). The acuteness of Wood's moral assessment is beside the point for present purposes; what matters is that the normative qualms with blatant, overt manipulation are clearly intelligible and relevant. But the covertness thesis cannot make sense of this idea at all. If the covertness were true, then such an assessment would exhibit conceptual inaptitude, a failure to understand what manipulation really is. But responding to Wood in such fashion – pointing out that what he finds morally abhorrent is not really manipulation – seems misguided precisely for the reason that there are

[9] As noted in the introduction, it is an interesting *further* question whether we can reasonably describe the Twitter algorithm as manipulative (in addition to, or instead of, the creators of the content). I address this question in (ref omitted for review).

cases of overt manipulation. Of course, one could claim that we are dealing with coercion here (following from the continuum thesis), but that does not seem plausible either.

These cases of manipulative interaction have already been discussed in the literature, though often not explicitly as a case against the covertness thesis.[10] It is important to explicitly address the covertness thesis in light of two considerations. First, insights that have long been gained in debates about normative ethics should also be considered in the increasing literature in digital ethics about (online) manipulation, where they are currently missing from the debate. Second, doing so helps to raise and then solve a neglected issue, the shifting of the conditions of manipulation from the manipulator to the manipulatee, as discussed below, which is a direct consequence of accepting the covertness thesis.

We can now turn to the problematic implications of the covertness thesis. The following fundamental theoretical problem with the covertness thesis is independent of the counterexamples presented above and the intuitions one may or may not have about them. The covertness thesis implies the following:

Covertness counterfactual: Had the manipulatee uncovered the hidden influence, it would not have been manipulation.

To illustrate, if a sneaky car salesperson tries to appeal to a buyer's sense of pride (e.g., 'imagine what your colleagues would think if they see you in that car!') to upsell the buyer – thus showing manipulative intent – and the buyer uncovers

[10] For critical discussion of the requirement that manipulation be covert, see also Gorin (2014, 74ff.); (Barnhill 2014, 59ff.), (Noggle 1996, 43–55), (Fischer 2017, 45ff.).

the attempted hidden influence, then there is no manipulation, according to the covertness thesis. But that counterfactual is both conceptually and morally problematic.

The conceptual problem is that the covertness counterfactual misallocates the grounding facts of manipulation from the agent (the manipulator) to the patient (the potential manipulatee). By 'grounding facts', I mean the natural facts that determine whether a given action counts as manipulation or not. Presumably, what makes a given action a case of manipulation is, amongst other things, what the actor does (e.g., Iago scheming, the car-salesman triggering, etc.) and *not* how the patient is *reacting* to the agent's action. But this is just what the covertness counterfactual implies: it says that we have a case of manipulation only if the manipulatee fails to respond in a certain way (namely, by uncovering the hidden influence) to the manipulator's action. To be clear: though it is, of course, true that the patient will to some extent determine whether manipulation happens (e.g. whether the patient is harmed by the manipulative action, or whether he or she consents to it), it is false that the existence of manipulation counterfactually depends on the patient's reactions to the agent's action.

The conceptual problem with the covertness thesis is further accentuated by luck in manipulation cases. The covertness counterfactual implausibly implies that a lucky uncovering of the manipulator's hidden influence would make the manipulation disappear. At the very least, the covertness thesis leaves open how to proceed with cases that turn out to be manipulation by sheer luck, i.e., where the manipulatee *almost* happened to uncover the hidden influence.

The conceptual problem leads on to a moral problem for the covertness thesis. The covertness counterfactual suggests the manipulatee might have the ability to prevent the manipulation from happening in cases in which he or she can uncover the influence. It is plausible that in such cases the manipulatee has an obligation to prevent the manipulation from happening. But that suggests that cases of manipulation can involve a *moral failing* by the victim of the manipulation. That, however, seems to be the wrong conclusion to draw and, therefore, the covertness counterfactual – and hence the covertness thesis – must be false.

Finally, note that there is a significant practical problem that follows from the covertness counterfactual. That is that policy recommendations might just too easily come down on the wrong side of things. The view demands either less covertness from would-be manipulators *or more* awareness from would-be manipulatees. Concrete policy proposals will be more likely to go the route of less resistance, and educating the public, rather than legislating the perpetrators, will often be more politically feasible. But it is undoubtedly the wrong conclusion to draw from noticing a large amount of manipulation that we should educate the public more, at least insofar as this is understood as potentially *completely eradicating* manipulation, as implied by the covertness thesis.

Proponents of the covertness thesis have two plausible replies. First, they could weaken their thesis as follows:

Covertness*: Manipulation is *attempted* hidden influence.

That concession would avoid the conceptual and the moral problem because the covertness counterfactual would no longer hold. In that respect, it seems a plausible option for proponents of the covertness thesis to take. However, some of

the proponents of the covertness thesis explicitly reject Covertness* (Susser, Roessler, and Nissenbaum 2019b). Also, that concession weakens support for the popular and influential view, discussed above, that *the actual increase in hidden influence* is a sign of manipulation. So, this is not a good option to take for proponents of the covertness thesis.

Second, proponents of the covertness thesis could distinguish between into manipulativeness and manipulated behaviour, and claim to talk about manipulated behaviour all along, which depends primarily on the patient. I am sympathetic to this response, and it is similar to some other recent accounts of manipulation in the literature (e.g. Klenk In press). But that is not a response on behalf of the covertness thesis because, again, the covertness thesis, if anything, is a view about manipulati*ve* behaviour, not manipulat*ed* behaviour. After all, that seems to be the reason why proponents of the covertness thesis often allude to the self-interested intentions of the creators of intelligent machines (e.g. Zuboff 2019).

Therefore, the covertness thesis is in deep trouble. It implies a conceptually problematic counterfactual which, in turn, has troublesome moral and practical implications. One could weaken the covertness thesis, but only at the cost of giving up its initial attractiveness. It would, therefore, seem like rejecting the covertness thesis and admitting that manipulation is only sometimes, but not necessarily, hidden is the only way out.

But we must not forget the demarcation dilemma. Rejecting the covertness thesis seems right from one perspective, but would throw us right onto one of the horns of the demarcation dilemma, and thus threaten us with abandoning manipulation as a clearly demarcated phenomenon altogether. Resolving this

tension will require an account of manipulation (or, to be precise, of at least one necessary condition of it) that avoids the pitfalls of the covertness thesis while at the same time showing how manipulation differs from coercion.

The main aim of this paper has been to raise this challenge and to bring it into sharper relief. The implications are clear for the field of digital ethics. An account of manipulation is sought that avoids the pitfalls of the covertness thesis without blurring the line to other types of social influence, notably coercion. In developing such an account, it will be interesting to see whether the seemingly striking feature of current machine-human influence – the subtlety and covertness of influence – can be left out of the definition of manipulation and yet one will find the currently dominant view of *increased* online manipulation vindicated. We will turn to such an account next.

3 Synthesis: Manipulation as careless influence

Let us briefly take stock. We have seen that the covertness thesis is appealing and plausible but that it succumbs to objections. The immediate problem that follows from the argument against the covertness thesis is that we have lost a way to distinguish manipulation from other forms of social influence, such as coercion. In this section, I introduce and defend an alternative account of manipulation and demonstrate how it helps to demarcate manipulation from coercion.

I defend the view that manipulation is careless influence. More precisely:

Manipulative action: M aims to manipulate a patient S if and only if

a) M aims for S to do, think, or feel b through some method m

and

b) M disregards whether m reveals eventually existing reasons for S to do, think, or feel for b to S.

Manipulative action thus understood requires the manipulator to intentionally employ some way of influencing the target to bring about a form of behaviour in the target, together with a lack of intention to reveal to the target any reasons that might exist to act in accordance with the manipulator's aims.

A typical manipulative action according to my account has the manipulator thinking roughly as follows: 'I want you to perform behaviour b, so I do action m, *and I would have chosen m even if it did not reveal your reasons for doing b to you*.' Naturally, this is but a gloss, and the italicised consideration need not be conscious to the manipulator. The manipulator intends his or her influence to have a particular effect on the manipulatee and chooses his influence accordingly, but he or she does not care whether the chosen means of influence reveal any reasons for exhibiting the intended behaviour to the manipulatee.[11]

Importantly, the necessary criterion for manipulative action is *not* that the manipulator is careless about whether the manipulatee has reasons to perform the desired action (e.g. the car-salesmen caring about whether the manipulatee has reasons to buy that car). Rather, the focus is on eventually existing reasons that are *potentially revealed through the manipulator's chosen means of influence.* We have a case of manipulation if and only if the manipulator does not care whether his or her means of influence reveals eventually existing reasons to the

[11] Note that the account lends itself to a counterfactual analysis – but not about the effectiveness of the influence (along the lines of 'had I not used manipulative method M, S would not have done p') but rather in terms of the propensity of the chosen method of influence to reveal reasons to the influenced person. See (MS ref omitted) for discussion. Thanks to [omitted] for comments on this point.

manipulatee. Therefore, the proposed account is quite distinct from influential but problematic, accounts of manipulation in terms of reason-undermining (e.g. Noggle 1996; Baron 2003; Beauchamp and Childress 2008; Blumenthal-Barby and Burroughs 2012).

Some manipulators like parents and liberal paternalist choice architects do care about a manipulatee's reasons for exhibiting the intended behaviour, but they do not care about revealing these reasons through their chosen method of influence. Parents may not care to appeal to the grounds of an action but instead opt for other means that maximise the likelihood of the action getting done because their children do not sufficiently grasp reasons yet (Kohlberg and Mayer 1972). Paternalist choice architects may not aim for increasing their subjects' insights because other means of influencing are more effective (Thaler and Sunstein 2009). In both cases, whether or not there are reasons for the manipulatee to act, and whether they are revealed through the chosen method, is a mere side-effect. The factor that turns both cases into cases of manipulative action is the fact that the agent in each case exercises *careless influence* – careless in the sense that the agent is not concerned with choosing a means of influence that may be reason-revealing to the patient.

The proposed account of manipulative action thus defends necessary and sufficient conditions for an action to count as manipulative. I will henceforth refer to the account as the MACI view, because it defines Manipulation As Careless Influence. Naturally, a moral question arises in response to manipulative action as defined by the MACI view, namely whether manipulation – or careless influence – is morally permissible. I suspect that it often is, as in cases where

parents interact with younger children, but I shall not pursue that question in what follows. It is important to keep in mind that the discussion of manipulation can proceed on the assumption that manipulation – in itself – is not a moralised or negatively connotated concept (cf. Coons and Weber 2014). The MACI account aims to capture what is distinct about manipulation. It captures well the intuitions recorded by philosophers engaged in conceptual analysis of the concept of manipulation.

Though this is not the place to demonstrate this through a case-by-case analysis, it is helpful to briefly contrast the MACI account with Cass Sunstein's account of manipulation, because there are important parallels and differences.[12] Sunstein suggests that an action is manipulative "to the extent that it fails to sufficiently engage or appeal to [the manipulatee's] capacity for reflection and deliberation" (Sunstein 2016b, 82, 2016a). Sunstein's account and the MACI account will often coincide (that is, provide classifications of a social influence as manipulative). When you are careless about how you influence someone (which counts as manipulative according to MACI), you may also fail to sufficiently engage or appeal to that person's capacity for reflection and deliberation (which counts as manipulative on Sunstein's account).

However, there is no material or substantive connection between both views, and I will highlight just two differences. First, Sunstein construes manipulation overly narrow in terms of a (sufficient) lack to engage deliberation and reflection. But many influences that do not at all aim to engage deliberation and reflection are *not* manipulative, contra Sunstein's account (cf. Barnhill 2016). MACI

[12] Thanks to an anonymous referee for emphasising the link to Sunstein's account.

recognises that we often have perfectly good reason for non-deliberative and unreflective thoughts, emotions, and behaviour. Failing to care about revealing those reasons is manipulative, not the fact that one does not aim to make them available for reflection and deliberation. Second, the MACI account is not just a broader (and more accurate) version of Sunstein's account, but it is fundamentally different. Sunstein's account relies on an implicit theory about appropriate influence to *identify* manipulation, whereas the MACI account allows us to identify manipulation in a non-moralised fashion and to use a theory about appropriate influence to *evaluate* the moral status of manipulation. The MACI account thus avoids a negatively connotated concept of manipulation and it still maintains the idea that the *wrongness* of manipulation can be a graded affair, even though an action just is or isn't manipulative.

With the distinction to Sunstein's account drawn, it should suffice to say that the five cases of manipulation recounted in the previous section, though vastly different, all turn out correctly to be cases of manipulation according to MACI. Note, however, that MACI does not tell us what it is *to be manipulated*. So, the MACI view can be completed as a view on manipulation only with a corresponding view about the nature of manipulat*ed* action. Existing discussions of manipulation rarely point out that there might be a difference between the two and yet most end up focusing on manipulative action almost exclusively. It is thus in continuation with existing work that the account at hand focuses on manipulative action, although I hasten to add that some commentators have been calling for a theory of manipulated action (Klenk and Hancock, J. 2019). With that caveat, we can

return to the question of how the MACI view resolves the tensions created by the reaction of the covertness thesis.

We should begin by noting that the MACI view saves an important intuition behind the covertness thesis. It saves the idea that there is something worrisome or problematic about being left in the dark on issues that concern one's life. Manipulation is intuitively worrying because it often does leave its victim in the dark about the bigger picture in which he or she acts. People seem generally uncomfortable with this idea, even if they get evidence that they would be better off as an effect of being left in the dark about aspects of their treatment (cf. Meyer et al. 2019). The covertness thesis can easily capture that intuition. But so can the MACI view. In contrast to the covertness thesis, however, MACI locates the problem not in the *de facto* unawareness of intentions or influences (which is not required for manipulation, as shown above), but in the intentional *lack of care* of the manipulator to help the victim to an insight into the bigger picture of his or her action.

The MACI view also evades the demarcation problem. It allows us to sharply demarcate manipulation form other forms of social influence by defining manipulation broadly. Manipulation, on this view, is a broad category that may well encompass instances of coercion. That is not a problem. Nobody has said *how* the demarcation should be done, only that it is needed. Coercion may thus be a form of manipulation, distinguished from manipulation by involving further components such as the use of force and the diminishing of agency.

The MACI view also overcomes the conceptual, moral, and practical problems raised for the covertness thesis discussed above. The MACI view overcomes the

conceptual problem because it makes manipulation counterfactually dependent on the actor's actions and intentions, but not those of the patient. This is exactly as desired, as per the earlier discussion. With the counterfactual rectified, the moral problem thus vanishes as well. There is also no practical problem.

Therefore, the MACI view allows us to successfully distinguish manipulation from other forms of social influence while at the same time avoiding the conceptual, moral, and practical pitfalls of the covertness thesis. That is a good result.

Before turning to a broader implication of the MACI view for digital ethics and specifically the debate about the ethics of digital behavioural technologies, it is prudent to address two potential objections to the MACI view. First, critics might object that the proposed account falls prey to the same problem that befalls the covertness thesis because the account makes manipulation depend entirely on the perpetrating agent. If that is the case, then we determine whether something is manipulative solely by looking at the perpetrator and, therefore, we violate an intuition about the two-sidedness of the manipulation relation. However, the objection misinterprets the criticism of the covertness thesis. The covertness thesis gave rise to a problematic counterfactual that implied that the manipulatee's own action instantiates the manipulation. That is not the case for the MACI view. Moreover, it should be noted that the practical problem above does not arise, because the MACI view asks us to focus on the perpetrators which is, arguably, the correct response to cases of manipulation.

Second, and more fundamentally, critics might object that the 'careless influence' idea appears to make manipulation not very demanding in the sense

that quite a lot of social influences may come out as manipulative. Whenever an actor does not care whether his or her chosen means of influence help the patient to see their reasons for action, then that action will count as manipulative. Notably, the MACI view invites us to view very many economic interactions as manipulative because the seller does not aim at revealing the buyer's reasons. However, this is not necessarily a problem for the view because the view does not imply that everything that is manipulative is also problematic (all things considered). As suggested above, we should resist the temptation to view manipulation as a negatively connotated concept from the start.

With the improved view on manipulation at hand, we can now look ahead at some broader implications of the MACI view for current digital ethics. One important implication concerns the reasons for taking seriously the possibility or threat of manipulation by or through digital behavioural technologies. As noted in the introduction, some scholars suggest that manipulation is a particularly problematic and widespread phenomenon in the online sphere. The MACI view may confirm that suspicion, though for reasons altogether different than those that have been offered so far. Recall that the covertness thesis seemed a formidable view partly because it could make sense of the intuition that there is something wrong with intelligent machines that subtly and covertly influence humans.

In light of the preceding argument, however, the alleged subtlety and covertness (insofar as it exists) is not per se a problem. Recalcitrant intuitions to the effect that there is indeed something wrong with these digital behavioural technologies, however, are not necessarily misguided. They may point to the fact

that intelligent machines influence us in a careless fashion, and perhaps even necessarily so, as Klenk (In press) has recently argued. If that view is along the right lines, then the problem is not that machines are influencing us in a particular way, but rather the (lack of) reason for why they are doing so.

Thus, the ethical problem with manipulation by or through digital behaviour technologies is *not* that technology offers even greater potential for hidden influence (though that may be true as well, especially for those with low technological literacy). Rather, the problem is that many of those influences may not be *aimed at revealing reasons to users*. These influences would qualify as manipulative, and they would not enhance people's ability to act in reason-responsive ways.

For the debate in digital ethics, this means that more attention should be paid to the *grounds* that determine which particular techno-social influences are applied to humans. For example, why are reminders and notifications (which are particular types of techno-social influences) used in applications such as health-care apps like 'Headspace' or social media platforms? Are they intended to increase engagement, no matter what? Or are they intended to show the users that there are indeed reasons to follow suit with the desired behaviour, e.g., do a mediation or check one's messages? The former case would count as manipulative, according to MACI, but the latter would not.

Of course, there are difficult methodological questions about how to determine the intentions behind a given techno-social influence to determine whether it is manipulative or not, and it can be shown that there are sound and satisfactory approaches (ref omitted for review). In short, as long as a method of

influence (e.g., a notification, a particular design, or salient picture) is chosen because it reveals reasons to the user, then the influence is reason-oriented and not manipulative. This can be assessed using counterfactual reasoning. For example, it can be asked whether the designers of the 'Headspace' app would have chosen 'Mindful Moments' push-messages had they not revealed reasons for being mindful to users. If the answer is 'no' (e.g., because they introduced that feature to maximise engagement with the app, for commercial reasons) then those influences are manipulative.

Once we turn our attention to the manipulativeness of digital behavioural technologies themselves (that is, when we consider them to be agents of manipulation, after having settled the relevant question outlined in the introduction), we can ask whether machines even have the capacity to aim to reveal reasons to human users. Since most artificially intelligent machines cannot even comprehend causality (Pearl and Mackenzie 2018), let alone the more intricate 'reason for' relation, there looms a principled obstacle for non-manipulative influence by machines.

Admittedly, much more has to be said about how to assess manipulation on the MACI account in concrete cases. For the purposes of this paper, however, the important result is that investigations of manipulation (in digital behaviour technologies) must look for (lack of) reason-orientation in the sense just sketched *rather than* the existence of hidden influences. A lack of reason-orientation should count for manipulation, and, therefore, raise further questions about appropriate moral and political reactions. Whether the influence is hidden or not, however, is not of primary interest.

4 Conclusion

Must social influence be covert to be considered manipulation? The question is important amid increased efforts to understand what manipulation is, especially in the light of advances in digital behavioural technologies. The covertness thesis implies that manipulation must be covert. However, the covertness thesis is false. There is overt manipulation, but it is always a form of *careless* influence. The paper has demonstrated the problems for the covertness thesis, both through counterexamples and by revealing a deeper theoretical problem with its implications for dealing with manipulation, and then discussed a superior alternative. It has shown that an account introduced at length elsewhere has superior features and is able to resolve the tensions around the carelessness account.

Though much of the argument has focused on the concept of manipulation, there are important links to the debate about the ethics of (digital) behavioural technologies. We can assess these technologies both in terms of their consequences or effects on human individuals and groups, but also in terms of the value of the nature of their influence on us. In order to enable such an analysis, which, it stands to reason, might reveal ethical problems and opportunities *specific* to digital behavioural technologies, we need an understanding of manipulation. We need that understanding to be able to classify whether a given influence counts as manipulative or not. We can then assess whether humans or organisations manipulate *via* digital behavioural technologies and whether those technologies may themselves do the manipulating. The account of manipulation as careless

influence has shown that the intention of actors to reveal reasons to the people they influence determines whether they act manipulatively or not.

Future work needs to extend the critical analysis begun here to further components of manipulation, eventually aiming to provide a full list of necessary and sufficient conditions. In particular, it seems promising to follow-up on the distinction between manipulated action and manipulative action, where very little has been done so far on the former concept, but which might prove to be of utmost utility in future philosophical, legal, and economic discussions of digital ethics.

References

Barnhill, Anne. 2014. “What Is Manipulation?” In Coons and Weber 2014.

Barnhill, Anne. 2016. “I’d Like to Teach the World to Think: Commercial Advertising and Manipulation.” *JMB* 1 (3-4): 307–28. https://doi.org/10.1561/107.00000020.

Baron, Marcia. 2003. “Manipulativeness.” *Proceedings and Addresses of the American Philosophical Association* 77 (2): 37. https://doi.org/10.2307/3219740.

Beauchamp, Tom L., and James F. Childress. 2008. *Principles of Biomedical Ethics*. 6. ed. New York, NY: Oxford University Press.

Blumenthal-Barby, J. S., and Hadley Burroughs. 2012. “Seeking Better Health Care Outcomes: The Ethics of Using the "Nudge"." *The American journal of bioethics* 12 (2): 1–10. https://doi.org/10.1080/15265161.2011.634481.

Burr, Christopher, and Nello Cristianini. 2019. “Can Machines Read Our Minds?” *Minds & Machines* 83 (5): 1098. https://doi.org/10.1007/s11023-019-09497-4.

Burr, Christopher, Nello Cristianini, and James Ladyman. 2018. “An Analysis of the Interaction Between Intelligent Software Agents and Human Users.” *Minds and machines* 28 (4): 735–74. https://doi.org/10.1007/s11023-018-9479-0.

Cohen, Shlomo. 2018. “Manipulation and Deception.” *Australasian Journal of Philosophy* 96 (3): 483–97. https://doi.org/10.1080/00048402.2017.1386692.

Coons, Christian, and Michael Weber. 2014. “Manipulation: Introduction.” In Coons and Weber 2014, 1–16.

Coons, Christian, and Michael Weber, eds. 2014. *Manipulation: Theory and Practice*. Oxford: Oxford University Press.

DeVito, Michael A., Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. "How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation." In *CHI 2018: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems : April 21-26, 2018, Montréal, QC, Canada / Sponsored by ACM SIGCHI*, edited by Regan Mandryk, Mark Hancock, Mark Perry, and Anna Cox, 1–12. New York, New York: The Association for Computing Machinery. Accessed September 07, 2020.

Fallis, Don. 2009. "What Is Lying?" *Journal of Philosophy*, 29–56.

Fischer, Alexander. 2017. *Manipulation: Zur Theorie Und Ethik Einer Form Der Beeinflussung*. Berlin: Suhrkamp.

Floridi, Luciano. 2014. *Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press USA.

Frischmann, Brett M., and Evan Selinger. 2018. *Re-Engineering Humanity*. Cambridge: Cambridge University Press.

Goodin, Robert E. 1980. *Manipulatory Politics*. New Haven CT: Yale University Press.

Gorin, Moti. 2014. "Towards a Theory of Interpersonal Manipulation." In Coons and Weber 2014, 73–97.

Klenk, Michael. In press. “Digital Well-Being and Manipulation Online.” In *Digital Well-Being*, edited by Christopher Burr and Luciano Floridi. Accessed November 17, 2019.

Klenk, Michael. 2020. “How Do Technological Artefacts Embody Moral Values?” *Philos. Technol.*, 1–20. https://doi.org/10.1007/s13347-020-00401-y.

Klenk, Michael, and Jeff Hancock. 2019. “Autonomy and Online Manipulation.” *Internet Policy Review*. https://policyreview.info/articles/news/autonomy-and-online-manipulation/1431. Accessed February 28, 2020.

Kohlberg, Lawrence, and Rochelle Mayer. 1972. “Development as the Aim of Education.” *Harvard Educational Review* 42 (4): 449–96. https://doi.org/10.17763/haer.42.4.kj6q8743r3j00j60.

Kramer, Adam D. I., Jamie E. Guillory, and Jeff Hancock. 2014. “Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (29): 8788–8970. https://doi.org/10.1073/pnas.1412469111.

Krstić, Vladimir, and Chantelle Saville. 2019. “Deception (Under Uncertainty) As a Kind of Manipulation.” *Australasian Journal of Philosophy* 97 (4): 830–35. https://doi.org/10.1080/00048402.2019.1604777.

Merriam-Webster. 2020. “"Manipulation".” Accessed February 27, 2020. https://www.merriam-webster.com/dictionary/manipulation.

Meyer, Michelle N., Patrick R. Heck, Geoffrey S. Holtzman, Stephen M. Anderson, William Cai, Duncan J. Watts, and Christopher F. Chabris. 2019. “Objecting to Experiments That Compare Two Unobjectionable Policies or

Treatments." *Proceedings of the National Academy of Sciences of the United States of America* 116 (22): 10723–28. https://doi.org/10.1073/pnas.1820701116.

Noggle, Robert. 1996. "Manipulative Actions: A Conceptual and Moral Analysis." *American Philosophical Quarterly* 33 (1): 43–55.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect.* New York: Basic Books.

Rudinow, Joel. 1978. "Manipulation." *Ethics* 88 (4): 338–47. https://doi.org/10.1086/292086.

Sunstein, Cass R. 2016a. "Fifty Shades of Manipulation." *Journal of Behavioural Marketing* 213. https://doi.org/10.2139/ssrn.2565892.

Sunstein, Cass R. 2016b. *The Ethics of Influence: Government in the Age of Behavioral Science.* Cambridge: Cambridge University Press.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2019a. "Online Manipulation: Hidden Influences in a Digital World." *Georgetown Law Technology Review* 4 (1): 1–45. Accessed February 27, 2020.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2019b. "Technology, Autonomy, and Manipulation." *Internet Policy Review* 8 (2): 1–22. https://doi.org/10.14763/2019.2.1410.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge [Electronic Resource]: Improving Decisions About Health, Wealth, and Happiness / Richard H. Thaler, Cass R. Sunstein.* Revised and expanded edition. New York: Penguin Books.

van Dijk, Teun A. 1998. *Ideology: A Multidisciplinary Approach*. Thousand Oaks, CA: Sage Publications.

Ware, Alan. 1981. "The Concept of Manipulation: Its Relation to Democracy and Power." *British Journal of Political Science* 11 (2): 163–81. http://www.jstor.org/stable/193581.

Wood, Allen W. 2014. "Coercion, Manipulation, Exploitation." In Coons and Weber 2014, 17–50.

Zarsky, Tal Z. 2019. "Privacy and Manipulation in the Digital Age." *Theoretical Inquiries* 20: 157–88. Accessed February 27, 2020.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York, NY: PublicAffairs.